

**DATA CLEANSING SOLUTIONS - Technical Brief**

**The Case for Data Cleansing**

Have you ever wondered why you receive calls from your long-distance telephone company asking you to switch to their service or receive promotional materials from your mortgage company for a home loan you already have? The harsh reality is that for many companies, data integrity is so poor that they have no idea who a significant number of their customers are. The underlying cause for this confusion is the inconsistency in customer identifying information, such as discrepancies in name, address, social security number, and other unique customer attributes. Ultimately, this lack of data integrity translates into lower customer satisfaction, wasted resources, and money needlessly spent.

To address these problems with an organization's data assets, AMULET Development Corp. provides professional data cleansing services based on leading-edge technology.

This technical brief presents an overview of the AMULET data cleansing process using advanced tools from Microsoft SQL Server 2005. Our process is divided into the following areas:

- ♦ Analysis and integration of data sources
- ♦ Perform address hygiene against the U.S. Postal Service file
- ♦ Perform "intelligent" data matching process
- ♦ Perform customer selection with predictive analytics

Microsoft  
**SQL Server 2005**

As a Microsoft Certified Partner firm, AMULET utilizes the latest relational database, OLAP, and data mining technologies available with SQL Server 2005.

Some of the technologies and tools AMULET uses for implementing data cleansing processes include:

- ♦ SQL Server Integration Services (SSIS)
- ♦ Fuzzy Lookups and Fuzzy Grouping
- ♦ Data Mining Prediction Query transformation using the Microsoft Decision Tree algorithm.



**Analysis of Data Sources**

The first step in a successful data hygiene process is to analyze all data sources contributing to data components to the process. In an ideal situation, the data source is an enterprise data warehouse, but often individual data marts and/or legacy systems also are contributors.

Through data source analysis, we're able to identify critical factors leading to data hygiene problems. Common conditions for such problems include: non-validated city/state fields, inconsistent customer names, different formats for street address, transposed data, and missing or incomplete data.

In the adjacent figure, we see an example of a customer data source with potentially inconsistent information because names were entered differently; "Cathy" Jones is really the same as "Kathy" Jones.

	FirstName	LastName	Gender
▶	Karl	Jones	M
	Cathy	Jones	F
	Kathryn	Jones	F
	Kathy	Jones	F
	Kent	Jones	M

**Data Hygiene**

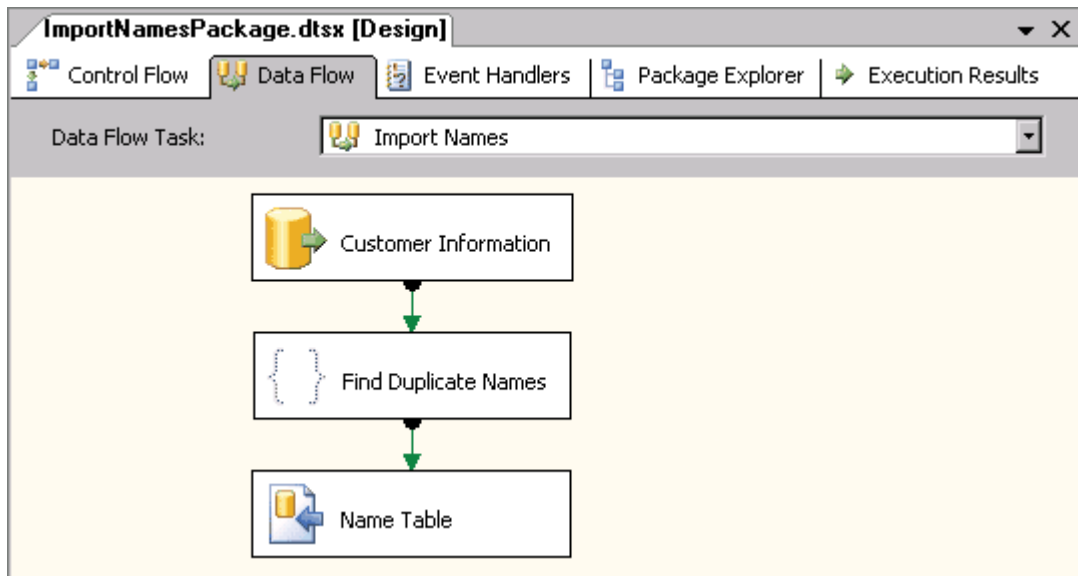
The data hygiene step cleanses and enhances existing address records in a batch mode. We compare each address in the data warehouse against the United States Postal Service ZIP+4 data file of addresses, plus address data from international postal authorities. The process will correct existing zip codes, add any missing ones and correct spelling errors. The resulting updated address information conforms to postal authority specifications for accuracy and consistency. Once addresses are cleansed and standardized, analysis and de-duplication of your data warehouse becomes easier and more reliable. The table shown below illustrates several data hygiene conditions and their remedies.

Before Hygiene	Action Taken	After Hygiene
1801 North Main Road, Lee, MA, 01238	Standardized to USPS address format, added Zip+4	1801 N Main Rd, Lee, MA, 01238-9063
36 Fletcher Avenue, Lexington, MA, 02421	Standardized to USPS address format, fixed incorrect Zip and adds Zip+4	36 Fletcher Ave, Lexington, MA, 02420- 3720
1900 Main St., Atlanta, GA, 30318	Address is flagged as unfixable because a post-directional is needed. There are 6 Main Streets in Atlanta, and batch doesn't know which one this is	Unfixable

## Intelligent Data Matching

The next step in the data cleansing process involves an intelligent data matching process that performs an accurate merge/purge operation using the *Fuzzy Lookups* and *Fuzzy Grouping* technology in SQL Server 2005. The process compares records and yields “similarity scores” that indicate which records likely represent the same entity and how strong that likelihood is. The calculated similarity score is based on the edit distance between a value and its potential match. The second metric that is produced is the “confidence value,” which indicates the amount of confidence the Fuzzy lookup algorithm has in the match that it found. The AMULET intelligence data matching process uses this number to determine whether to accept the value the algorithm found or manually check the value.

AMULET design and develops custom SQL Server Integration Services packages such as the one depicted in the figure below to satisfy your specific data matching needs. SSIS also offers the ability to embed custom .NET programming code to implement special business rules to guide the data matching process.



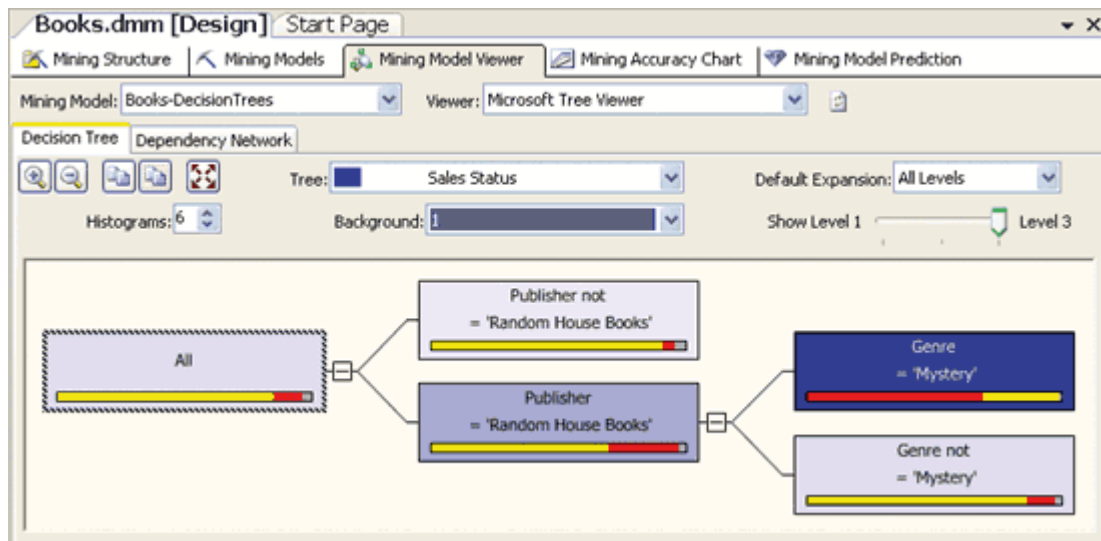
The figure below shows the results of an intelligent data matching process. Notice the new fields generated: CleanFirstName, and CleanLastName. The SimilarityScore value can be used to drive a subsequent automatic merge or manual verification process.

Key_In	OriginalImportFirstName	OriginalImportLastName	Gender	Key_Out	CleanFirstName	CleanLastName	SimilarityScore
1	Karl	Jones	M	1	Karl	Jones	1
4	Kathy	Jones	F	4	Kathy	Jones	1
2	Cathy	Jones	F	4	Kathy	Jones	0.824352920055...
3	Kathryn	Jones	F	4	Kathy	Jones	0.749075591564...
5	Kent	Jones	M	5	Kent	Jones	1

## Predictive Analytics

Once the data cleansing operation is complete, much benefit can be gained from using data mining techniques to provide predictive analytics for maximum use of your data assets. The accuracy of predictive analytics increases significantly when used on cleansed data sources. Predictive analytics, using general purpose data mining algorithms, provide the means to find patterns and make predictions within a set of data.

As an example, AMULET BI specialists can use a data mining algorithm to examine your customer's buying habits and determine what goods or services a particular customer is likely to buy next or predict which of your customers might be likely to shop around for a different supplier. For a retail business such as a bookstore that gets a list of available books each week from a distributor, there is an inventory cost associated with each book kept in stock. A good approach is only to stock books that are likely to be good sellers. The SQL Server *Data Mining Prediction Query* transformation task is able to make these predictions. The *Decision Trees* algorithm is a classification type of algorithm that works well for predictive modeling such as determining sales success status. The figure below shows the SQL Server Mining Model Viewer with the decision trees results predicting that "Mystery" titles from the "Random House Books" publisher have the highest likelihood of sales success.



## Who is AMULET Development Corp?

AMULET Development Corp. is a business intelligence solutions firm founded in 1995 to provide quality technology solutions for businesses in a broad range of industries. Specializing in Microsoft server, Web, and database technologies, we've developed many high profile e-business software applications. Our current focus is business intelligence, analytics, and data mining using contemporary technology to help enterprises better utilize valuable data assets.

## For information on AMULET's Data Cleansing services please contact us:

Sonya Franklin 1-877-722-7393

info@amuletc.com

www.amuletc.com

